DHQ: Digital Humanities Quarterly

2018
Volume 12 Number 2

Predicting the Past

Tobias Blanke <tobias_dot_blanke_at_kcl_dot_ac_dot_uk>, King's College London, Department of Digital Humanities

Abstract

Digital humanities have a long tradition of using advanced computational techniques and machine learning to aid humanistic enquiry. In this paper, we concentrate on a specific subfield of machine learning called predictive analytics and its use in digital humanities. Predictive analytics has evolved from descriptive analytics, which creates summaries of data, while predictive analytics predicts relationships within the data that also help to explain new data. Predictive analytics uses machine learning techniques but also traditional statistical methods. It uses properties (or features) of the data to predict another target feature in the data. Machine learning is used by predictive analytics to establish the rules that given a certain combination of features make the target more or less likely. Predictive analytics can thus be considered to be a technique to machine-read data. The paper discusses the background of predictive analytics, its use for predicting the past and finally presents a case study in predicting past gender relations in a historical dataset. Predicting the past is introduced as a method to explore relationships in past data.

Introduction

Digital humanities have a long tradition of using advanced computational techniques and machine learning to aid humanistic enquiry [Juola 2008] [Jockers and Witten 2010] [Röhle 2012]. [Blanke and Hedges 2013] have shown how machine learning has become part of the scholarly infrastructure of digital humanities while [Anderson et al. 2010] place it firmly within the methodology commons of digital humanities. Beyond machine learning's application in academic research in the digital humanities, several initiatives have explored the relationship of machine learning to cultural heritage institutions. The British Library, for instance, has developed a machine learning experiment [British Library 2015] to engage computer science students with digital heritage collections. While this experiment works with images, most digital humanities and machine learning approaches use textual data [Argamon and Olsen 2009], exploiting text mining techniques such as clustering or topic modelling. 1

With recent advances, machine learning has led to several new application areas. In this paper, we concentrate on predictive analytics for digital humanities, as it has evolved from descriptive analytics [Abbott 2014]. Descriptive analytics creates summaries of data, while predictive analytics predicts relationships within data that also help to explain new data. Predictive analytics employs both machine learning techniques and traditional statistical methods. It uses properties (or features) of the data to predict another target feature in the data. Machine learning helps establish the rules that given a certain combination of features make the target more or less likely. Predictive analytics can thus be considered to be a technique to machine-read data [Abbott 2014]. 2

In the perception of the public, predictive analytics has been linked to predicting the future [Anadiotis 2016]. However, the academic literature has a more varied view on predictive analytics. The ideas of predicting the future are expanded to the art of the predicting the present [Choi and Varian 2012] and also the "past" [Schutt and O'Neil 2013]. Google's chief economist, Hal Varian, contends that queries that users enter into Google's search engine describe how they feel and act in "real time" [Choi and Varian 2012]. Alongside predicting the future and the present, "predicting the past" is the maybe counter-intuitive prediction of past performance by joining historical datasets. In criminology, for instance, predicting the past could lead to "more precise attribution of past crimes, and the apprehension of suspects" [Wang et al. 2013, 515]. 3

Predicting the past creates new relations between datasets rather than anticipating the future. In this paper, we discuss predictive analytics as a method to understand relations of the past and proceed by exploring the prediction of genders in an example dataset. Predicting the past is introduced as a method examining datasets about the past using machine learning. 4

Background and Methodology

5 Computational prediction has attracted much interest recently as part of the new field called “predictive analytics”. Abbott is one of the most famous practitioners in the field [Abbott 2014]. For him, predictive analytics work on “discovering interesting and meaningful patterns in data. It draws from several related disciplines, some of which have been used to discover patterns in data for more than 100 years, including pattern recognition, statistics, machine learning, artificial intelligence, and data mining.” [Abbott 2014, 29]. Predictive analytics brings together all these traditions, but is commonly misunderstood and reduced to attempts to predicting the future. Fear and critique of computational prediction is often linked to a dystopian future where the movie *Minority Report* comes to life and all our existence is pre-determined and governed by analytics. However, predictive analytics, as this paper will show, is a much more diversified discipline and offers digital humanities researchers exciting opportunities to read their data.

6 Predictive analytics is about developing meaningful relationships in any data, machine-reading it and making sense of it. Compared to traditional analytics, predictive analytics is driven by the data under observation rather than primarily by human assumptions of the data. It strives to automate modelling and finding patterns as far as this is possible. But in principle “predictive analytics doesn’t do anything that any analyst couldn’t accomplish with pencil and paper or a spreadsheet if given enough time; (…).” [Abbott 2014, 30].

7 Most attention in predictive analytics has been given to predicting the future and present. Schutt and O'Neil, however, also offer a new idea of predicting the past [Schutt and O'Neil 2013]. In 2007, new large amounts of electronic health records (EHR) and related data allowed the Federal Drug Agency (FDA) to set up new monitoring programmes for drugs. Among the new programmes, the Sentinel project worked on large-scale data integration. For Schutt and O'Neil, these integrated datasets were the foundation of novel research attempts to predict the past. They cite the “Observational Medical Outcomes Partnership (OMOP)” in the US that investigates how good we are at predicting what we already know about drug performance in health using past datasets. Once OMOP had integrated data from heterogeneous sources, it began to consider predicting the past of old drug cases and how effective their treatments were.

8 In this sense, predicting the past tries to understand how “well the current methods do on predicting things we actually already know” [Schutt and O'Neil 2013]. Rather than targeting new knowledge on the past directly, predicting the past is here understood as a means of evaluating existing methods. An existing example that implies predicting past events by joining historical datasets, is the identification of historical spatio-temporal patterns of IED usage by the Provisional Irish Republican Army during “The Troubles”, used to attribute “historical behaviour of terrorism” [Tench et al. 2016] . Analysing predictive policing, Aradau and Blanke demonstrate how past joined-up datasets are already used to develop profiles of suspects [Aradau and Blanke 2016]. The authors also present a methodology how to criticise predictions techniques and their abstracted similarities. Katz, Bommarito et al. offer a “generalized, and fully predictive model of Supreme Court voting behavior”’ in the United States [Katz et al. 2014].

9 For the study of the past, Weisskopf states the general interest in predicting the past: “There is also such a thing as a prediction of the past, "retrodiction," when conclusions are drawn with regard to the yet unexplored previous history of a given set of objects.” [Weisskopf 1984]. There remain, however, concerns in the humanities with regard to the predictability of the human condition. Firat captures this sentiment, when he questions the value of these approaches for understanding the nuances humanities research is interested in: “This is exactly why [statistical] approaches of this kind are antithetical to both explanation and understanding. They can predict, but since the techniques used lack information on or evidence of underlying characteristics, they cannot provide the answer to why the central tendency in one variable is related to the central tendency in another.” [Firat 1987]. Popper’s rejection of prediction in the human sciences as pseudo-science [Popper 1974] is often referred to in this context in order to underline the impossibility of prediction in the humanities. The study of cultural phenomena is too complex and different to be predicted in terms of statistical aggregates. Digital humanities are able to avoid such criticisms and engage with prediction by staying close to datasets, as we will demonstrate. Then, prediction does not have to be about grand explanations of history that Popper’s critique targets. It rather is about the distant reading of these datasets.

10 Digital humanities have started to engage with predictive analytics as a method of reading past data rather than as a way to find regularities and laws in human history. Manovich, for instance, explains the theoretical and conceptual background of predictive analytics as well as its applications in art history [Manovich 2015]. Several archaeological studies have used predictive techniques to explore datasets about ancient societies [Gaffney 2008] [Stanley 2012]. However, these studies are often focussed on archaeological data only and remain theoretical investigations [Jasinski 2017]. Lincoln goes furthest in investigating predictive analytics for the humanities and presents a case study on predicting the past in the context of art history [Lincoln 2017]. He is a data scientist working for Getty and uses machine

learning to “distil evidence into a cogent argument. […]. Like an X-ray, a simplified model [in predictive analytics, TB] does flatten our view of the subject. But it also reveals a perspective that would otherwise have gone unseen” [Lincoln 2017]. Working for Getty, Lincoln’s predictive analysis investigates historical patterns of artwork sales rather than understanding the composition of scholarly concepts. Nevertheless, he provides a good overview of the steps and benefits involved in the use of predictive analytics.

This paper brings together Lincoln’s and others ideas about predicting the past in digital humanities with the suggestions of Schutt and O'Neil to use the approach to test existing models with past datasets [Schutt and O'Neil 2013]. We are concerned both with the exploratory power of models to understand past relations in data and with the methodological innovation and evaluation of existing models. Thus, we compare a traditional model of predicting past data using dictionaries with machine learning strategies. First, a classification algorithm is discussed and then three different rule-based learners are introduced. We can demonstrate how these rule-based learners are an effective alternative to the dictionary-based method and partly outperform it. As they perform at least as well as the traditional method, we propose that the machine learning approaches can be an alternative for digital humanities. 11

The second objective of this paper is to introduce the “predicting the past” methodology by means of an example dataset to predict gender relationships. We demonstrate how predicting the past can complement and enhance existing work in digital humanities. Any predictive analytics model works from available observations to predict so-called target values. These targets need to be extracted from the existing dataset. Labels are features to be predicted based on features that are chosen as predictors. Once we have a target, we need to understand which parts of the observations or features can be used to derive the target labels. Often the features have to be modified to improve the predictive performance. We will, for instance, merge various time-based features for our case study. This feature extraction and engineering is a very labour-intensive process. 12

The most important consideration in predictive analytics is to derive a model that generalises to new cases [Bowles 2015, 59]. Our aim is to predict past data. Thus, we do not have any new data but set aside data from the existing set to test how the prediction performs. We will indicate exactly which proportion of the data we hold out to test the performance of the model. In addition, we employ “cross-validation” to avoid overfitting the data, one of the main risks in predictive models. An ‘overfitting’ model is one that models existing training data too closely, which negatively impacts its ability to generalize to new cases. Cross-validation uses different partitions of the data into training and test data and creates a set of subsamples and thus helps avoid overfitting [Schutt and O'Neil 2013]. The training subset uses all-but-one subsample, while the model is validated on the remaining subsample that functions as a test subset. The final performance is then based on the average performance of all these different training and test partitions. 13

Overall, predicting the past borrows from predictive analytics the following first steps: 14

1. Engineering the necessary features and predictors
2. Defining the training targets
3. Training a model
4. Evaluating the model performance with test data to eliminate underperforming models.

However, predicting the past is also different from other predictive analytics approaches, as the model is not prepared to understand new observations but to analyse the relationships between existing ones, as outlined by Schutt and O'Neil. The aim is to understand which (minimal) set of features makes it likely that observation x includes target feature y. In our, case we would like to understand which combination of features make it likely that a historical person is of gender female, male or unknown. 15

The next step in our methodology is therefore to apply the best performing models to the whole dataset again rather than to the test data only, as it is standard practice in predictive analytics. The model then reveals the rules that govern the determination of gender, but also cases that should interest us further such as misclassifications or disagreements. This analysis is the main objective of our method. For instance, we investigate in detail which features contribute most to the performance of our prediction. Does an English location, for instance, make it more likely that the target gender is female or male or unknown? An increase in performance of the prediction indicates that the added features contain useful information to determine gender. Finally and following Schutt and O'Neil, we compare a number of prediction methods to find the one that is most reliable and useful for our analysis [Schutt and O'Neil 2013]. Predicting the past can thus be understood as a “spam filter” for various methods that could be used to predict historical categories in a dataset Schutt and O'Neil. Once we have achieved the optimal models even prediction errors will help us understand the relationships in the underlying data, as we will describe. These errors can be as significant as successful classifications for understanding relationships in the data. 16

Exploring the data and creating the prediction background

17 Before we can proceed with the predictive analytics, we need to unlock the data to get an understanding of how we can work with it. In particular, we work through a series of exploratory steps to prepare the predictive analytics. We will first get an overview of the cases and observations as well as the features that are used to describe each observation. Here, it is particularly important that we relate the number of observations to the number of features as only with the right amount of observations we have enough combinations of features that will allow us to build the right predictive model. Models need to have far fewer features (also called degrees of freedom) than observations. Secondly, we determine how many categorical and numerical features we will work with. Thirdly, we develop a strategy how to work with missing values, which are very common in historical datasets. Finally and in order to understand a possible direction of travel for the predictive analytics, summary statistics and plots are deployed.

18 In order to experiment with predictive analytics and demonstrate our approach, we decided on modelling gender predictions following on to the pioneering work by Blevins and Mullen [Blevins and Mullen 2015]. We try to understand which algorithmic gender prediction technique could identify genders best. Please note that we do not make claims about the possibilities to predict the actual genders but the genders as expressed in the data. Any underlying gender issues in the datasets will be reproduced and at no point do we claim that our algorithm understands gender better than what is in the data. As described, predicting the past does not mean to develop meta-theory of the past but to closely describe the data.

19 In order to facilitate reuse of datasets and future integration, humanities data journals have started to appear that offer an opportunity to publish and receive recognition for research data outputs. One of the first humanities data journals in Europe is the *Journal of Open Humanities Data*[1] published by Ubiquity Press. The first dataset published by the journal was “Vagrant Lives: 14,789 Vagrants Processed by the County of Middlesex, 1777–1786” by Adam Crymble, Louise Falcini and Tim Hitchcock [Crymble et al. 2015]. We have chosen “Vagrant Lives” as an example dataset, as it is currently the best we can expect from data research output in data journals. It contains the well-curated records of vagrant removals in the 18th century. At the time of writing, it includes 14,789 records about vagrant incidences that happened between 1777 and 1786.

20 The dataset was produced as part of the “Vagrant Lives” project [Crymble et al. 2015] and captures the work of Henry Adams, the vagrant contractor for the UK county of Middlesex: “Eight times per year at each Middlesex Session of the Peace, Adams submitted lists of moved vagrants to receive remuneration” [Crymble et al. 2015]. Parts of these records are, however, missing, as it is typical for historical datasets [Anderson et al. 2010]. Rather than leaving missing records empty, the creators of the dataset have helpfully indicated that the records are missing from the original rather than being left out in the transcription by adding specific values for the missing records. The entry “[unknown]” designates a missing record for the vagrants. Introducing an extra entry value for unknown records, should be seen as good practice for humanities data, as otherwise we do not know whether the records were incomplete in the first place. A predicting the past approach should reproduce these unknowns, as they correspond to the data and not outside background knowledge such as dictionaries of first names that determine their gender.

21 Next we explore the data using the Python programming language and in particular the PANDAS environment and MatPlot toolkit [Raschka 2015]. A typical entry in the dataset looks like:

```
df.iloc[0]
```

Vagrant ID Number	6625.1.1
Given Names	Mitchell
Surname	Bruce
Gender of Lead Vagrant	M
Relationship to Lead Vagrant	[lead vagrant]
Number of People in Group	1
Person Type	Solo Male
Vagrant Category	City Vagrant
Session Start Day	12
Session Start Month	9
Session Start Year	1784
Session End Day	14
Session End Month	10
Session End Year	1784
Session # (out of 8 annually)	6
URL of Primary Source	http://hri.shef.ac.uk/san/pl/SM/PS/LMSMPS50787...
Magistrate Name	John Hart
Taken From	House
Conveyed To	Cheshunt
Georeference (Taken From)	51.5321;-0.1066
Georeference (Conveyed To)	51.699888;-0.028486
Settlement (Micro Level)	Pomona
Settlement (Area Level)	[n/a]
Settlement County	Zetland
Settlement Country	Scotland
Settlement Georeference (Micro Level)	59;-3.25
Settlement Georeference (Area Level)	[n/a]
Settlement Georeference (County)	60.33333;-1.33333
Settlement Georeference (Country)	56;-4

Name: 0, dtype: object

Figure 1. Example Dataset in the PANDAS environment using the MatPlot toolkit

As we can see, the dataset does not contain a direct gender feature, but the gender of the vagrant is relatively easy to derive. The “Gender of Lead Vagrant” is recorded as is the “Relationship to the Lead Vagrant”. The latter is a categorical feature containing a limited number of different types of entries. If it states “[lead vagrant]” we can presume that the gender is identical to “Gender of Lead Vagrant”. Otherwise, the “Relationship to the Lead Vagrant” can be family relationships such as “wife”, “son”, etc., which again allow us to derive the gender. In the few remaining cases (often undefined children), we cannot conclude anything and declare the gender to be “unknown”. The column “Person Type” contains at first sight promising knowledge, too. It has values such as “Solo Male” or “Solo Female”. But it also contains not useful entries such as “Dependant” or “Group Leader”. So, we decided to rely on “Gender of Lead Vagrant” and “Relationship to the Lead Vagrant” to derive the gender but will use “Person Type” later to check whether our gender derivations were correct. We arrived at the following gender distribution, which we added to the vagrant dataset: 22

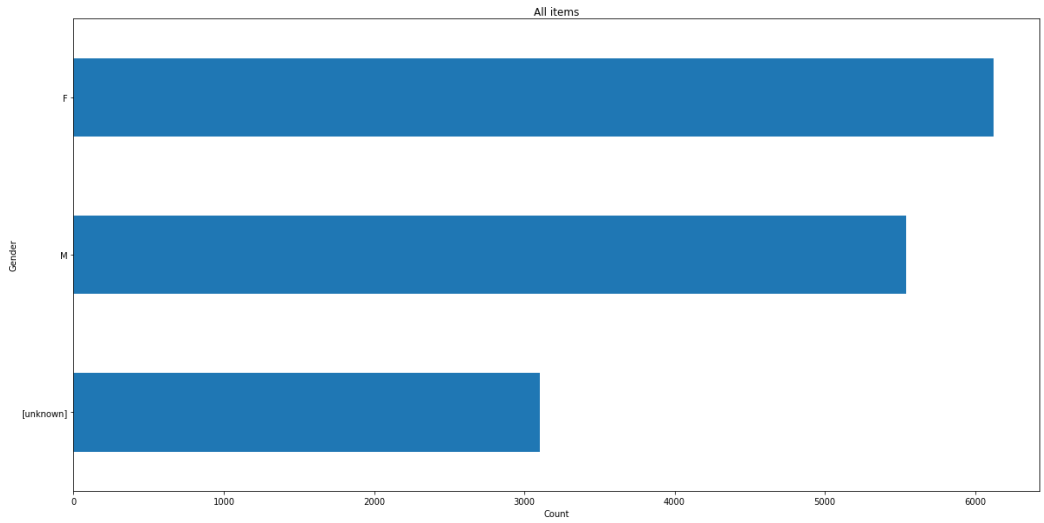

Figure 2. Genders in Vagrant Lives

In order to confirm that gender can be derived from the two features gender of the lead vagrant and relationship to the lead vagrant, we compare the feature “Person Type” with the newly derived gender feature. Table 1 shows that dependents dominate the list of unknown genders, while there are two single females that are recorded as having unknown gender. Apart from these two records, however, the dataset looks consistent. We have shown that our derivation of genders from two features is consistent with a third feature “Person Type” and that our target feature is consistent with the data. 23

Gender	F	M	[unknown]	All
Person Type				
Dependent	687	16	3050	3753
Group Leader	1358	755	9	2122
Single Female	4080	0	2	4082
Solo Male	0	4770	0	4770
Unknown Solo	0	0	43	43
All	6125	5541	3104	14770

Table 1. Table 1: Gender and Person Type in Vagrant Lives

Having gained an overview of the dataset's features and having defined the target feature gender, we can now start controlling the predictors and target features. In predictive analytics, predictors are those features that a model uses to predict a target feature, which is in our case the gender. The dataset is relatively large for a historical dataset with over 14,000 entries. Despite this large amount of data, prediction is made difficult by the fact that the data contains a lot features (29 in total). The prediction space is thus sparse, which is an obstacle to predictive analytics, because it is based on distances of data points in the feature space [Aradau and Blanke 2016]. "A]s the number of inputs increases, the number of data points needed to fill the space comparably increases exponentially." [Abbott 2014, 153]. With increasing numbers of features, the prediction performance first increases but then rapidly decreases. 24

Our feature density is clearly not high enough. As the reader might not be familiar with the consequences, let us discuss briefly a simple example of 2 features having on average 5 different values each. This means we have 5x5 or 25 individual possible data points. Let us furthermore expand our existing vagrant dataset a bit and assume we have 15,000 observations. With 15,000/25=600 we have more than enough data (600) to cover every single possible case. With 29 features, however, we need 1.86e+20(=5 to the power of 29) observations to cover all possible data points. This calculation is additionally based on the simplified assumption that all 29 features have 5 different values each, while many features in the Vagrant dataset contain more. Because most prediction work is grounded in the reorganisation of these kind of data spaces and thinking about how to bring the various observations together [Aradau and Blanke 2016], historical predictions are generally challenged by sparse data. To organise the predictor features, the number of overall features has to be reduced. Klingenstein, Hitchcock et al., for instance, demonstrate how this so-called "curse of dimensionality" can be addressed in text analytics by using a historical thesaurus to limit the vocabulary [Klingenstein et al. 2014]. 25

We must reduce the feature set and transform feature values into entities a prediction model can work with. In our dataset, we can first remove all the geo-references, as they are redundantly encoded in other columns such as "Taken from" and "Conveyed To". On the other hand, because we would like to explore temporal relationships, we keep all the relevant time features. But we can also reduce features here and merge all the Session Start and Session End columns into one POSIX date column. The Portable Operating System Interface (POSIX) is a way to standardize dates. Doing so, we not only reduce the number of features but also enable a whole set of data exploration techniques. Figure 3, for instance, shows when new magistrates first appear in the dataset. As the Figure clearly includes various local maxima and is overall skewed to the right, the data is unevenly composed, either because of a relevant trend we should investigate further or, as it is more likely in this case, because records contain missing values for the earlier dates. Finally, we transformed the newly created POSIX date features 'Session Start' and 'Session End' into absolute differences in seconds from a cut-off date. This quantification will help with our machine-learning approaches. 26

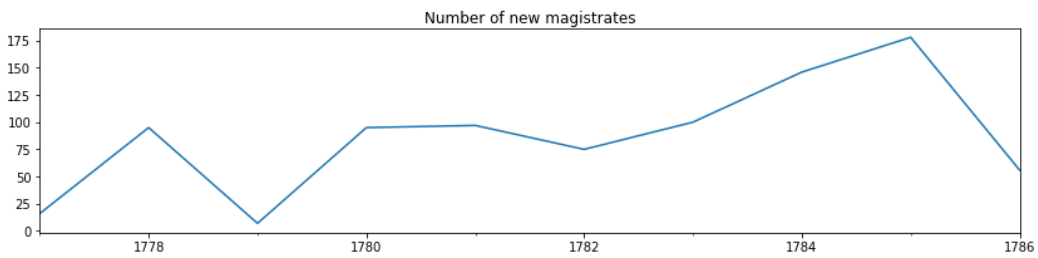

Figure 3. New magistrates in Vagrant Lives

For the rest of the features we factorized (value-indexed) the 9 categorical features and added them to the remaining 3 numerical columns that measure temporal relationships. In total, we therefore considered 12 features instead of 29; still 27

too many for the size of the dataset but a much better foundation. Later on, we will work on the further reduction by using feature selection techniques and will be able to reduce the overall number of significant features to 5.

As described above, our exploratory analysis concludes with summary plots of the data in order to better understand potential relationships the predictive analysis could target. Typical graphical techniques explore the connection between two features and provide new viewpoints on the data. A grouping by the “Taken From” feature, for instance, reveals that most vagrants were taken from their house without further location specification. This has twice the number of entries compared to the next place “Clerkenwell”. 28

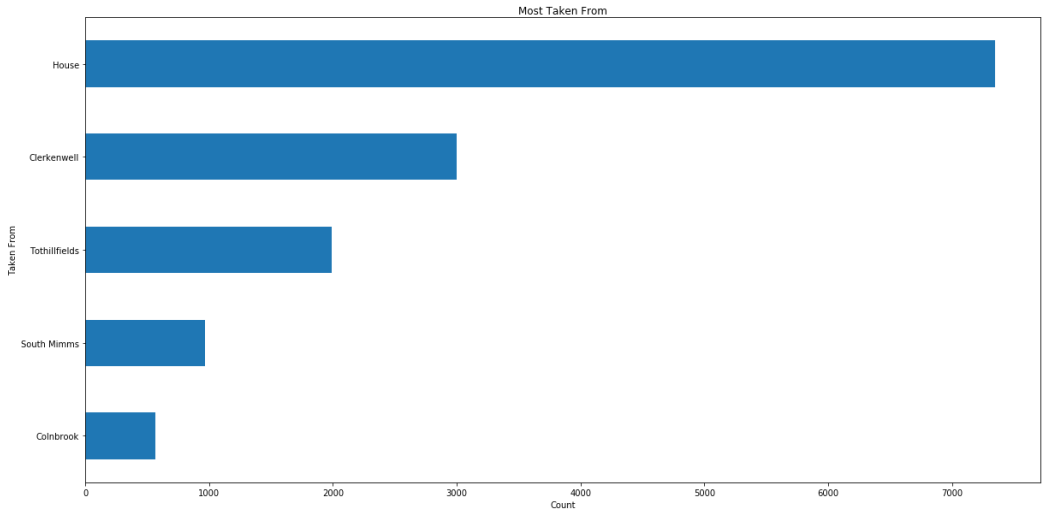

Figure 4. Vagrant locations

Similarly, we can determine that the most active magistrate is Richard Clark. 29

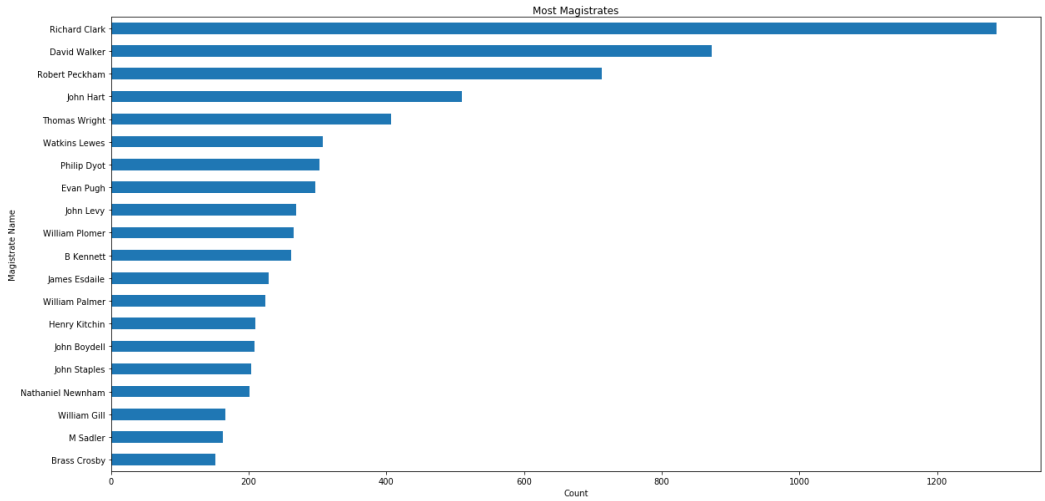

Figure 5. Case count for magistrates

However, data exploration has clear limitations when it comes to a larger number of (non-numerical) features, as they are common in the digital humanities. An exploration of two features and their relationship fits easily with the two-dimensional layout of paper, but for three features we already need a combination of graphs and axes to make the visualisation work. Figure 6 shows how three perspectives on the timeline of events can be integrated. We can identify the uneven distribution of the data, but also that most cases were recorded in the second half of the year. Beyond three features the visual representation becomes more and more complicated in an two-dimensional layout. Predictive analytics allows for a higher number of features to be integrated than possible for a typical exploratory analysis that targets plots. 30

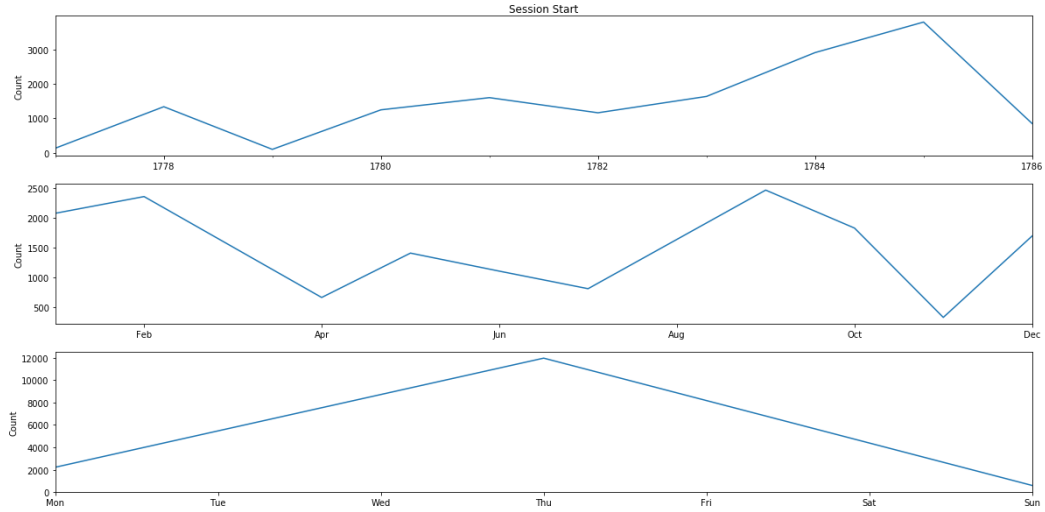

Figure 6. Three-dimensional comparison of session times

Finally, it is difficult to represent in plots the non-numerical, categorical data that dominates the humanities. The mosaic plot [Schutt and O'Neil 2013] is an example that is popular to present the relationship between categorical features, but its exact proportions are often difficult to decipher. It is, therefore, no wonder that visualisations in the digital humanities commonly exploit continuous features such as timelines or word counts [Jänicke et al. 2015]. Predicting the past on the other hand would be able to exploit methods that could work with categorical data more effectively than visualisations, but that also often means we work with complex and therefore opaque models, which need to be analysed carefully [Aradau and Blanke 2016]. 31

Shmueli asks whether “To Explain or to Predict?"' and presents both as necessary for developing insights but from a different viewpoint on the underlying data. For data exploration, “the role of the theory is very strong and the reliance on data and statistical modelling are strictly through the lens of the theoretical model.” [Shmueli 2010, 290]. He defines “explanatory model[l]ing as the use of statistical models for testing causal explanations.” [Shmueli 2010, 290]. Typical operations for explanatory modelling include statistical tests to ensure that the observations support an assumed theoretical relationship – often expressed as a curve [Schutt and O'Neil 2013]. Furthermore, where the theory is not directly observable, observations are linked to theoretical concepts. This step includes the use of established theories and knowledge to link concepts and observations. If exploration is about the further development of existing theories, a model is predictive and about new insights and generalisations from known insights to new ones. According to Abbott, prediction is about systematically determining the relationship between features beyond data visualisation and is itself embedded in iterative improvements of the developed model, which provide new insights at each step [Abbott 2014]. 32

In the following sections, we explore the predicting the past approach further by comparing different models to predict genders in Vagrant Lives. The first one uses a “theory” and history of first names and their relationships with genders. This approach is common in the digital humanities and can be seen as a baseline to compare other predictive analytics approaches. All these methods can help on the one hand side explain how historical decisions are divided according to genders and on the other hand side they have a very practical application as the imputation of missing values in historical datasets. As imputation is trying to derive missing values, it can be considered to be a (simple) prediction technique. 33

Predicting the past using ‘theories’

Blevins and Mullen provide an explanation as to why digital humanities should be interested in predicting genders. Gender values are often missing from datasets and need to be imputed [Blevins and Mullen 2015]. As humanities data will contain many empty records [Anderson and Blanke 2012], understanding imputation is fundamental to digital humanities. In this section, we will compare two methods based on the ideas in [Blevins and Mullen 2015] to impute missing values of genders. Both presented methods assign a gender to a vagrant using the given name. The first one uses a single dictionary of names with many international names, while the second one employs a dictionary, which is split according to time periods but always from English-speaking communities. Blevins and Mullen present an ensemble approach that is based on collecting many given names' dictionaries into an R script, which takes a vote count of all the different dictionaries on whether a given name is more likely to be male or female [Blevins and Mullen 2015]. 34

35

Dictionaries are used in many prediction scenarios and help not just to read first names but also determine sentiments [Grimmer and Stewart 2013] – among other things. For gender predictions using dictionaries, Blevins and Mullen discuss the many issues linked to such an approach [Blevins and Mullen 2015]. Firstly, names might change over time their gender. Leslie, for instance, changed from being a male name to a mainly a female name in the English-speaking world. There are secondly, however, even more serious issues with dictionary-based prediction, which could heavily influence our understanding of the underlying data. Dictionary approaches assume that the data is completely missing at random and that we are speaking about missing or wrongly assigned data and want to impute it with the correct data. The algorithm then determines the "correct" value of the gender by guessing it from the most common association with the name. But data can also be missing not at random, because, for instance, the way the data was gathered had issues. Missing out on genders can be a simple mistake but might also point to the impossibility to determine — in this case — gender at the point of gathering the data. Maybe the question of gender was simply impossible to answer for a certain vagrant, which would probably also be the most interesting cases. Recognising this, the creators of the vagrant dataset have included the class [unknown] for gender. It is the assumption of this article that a critical (computer) reading of data would require us to keep being aware of the critical distinction of why values are missing and not equate to "incorrect" entries.

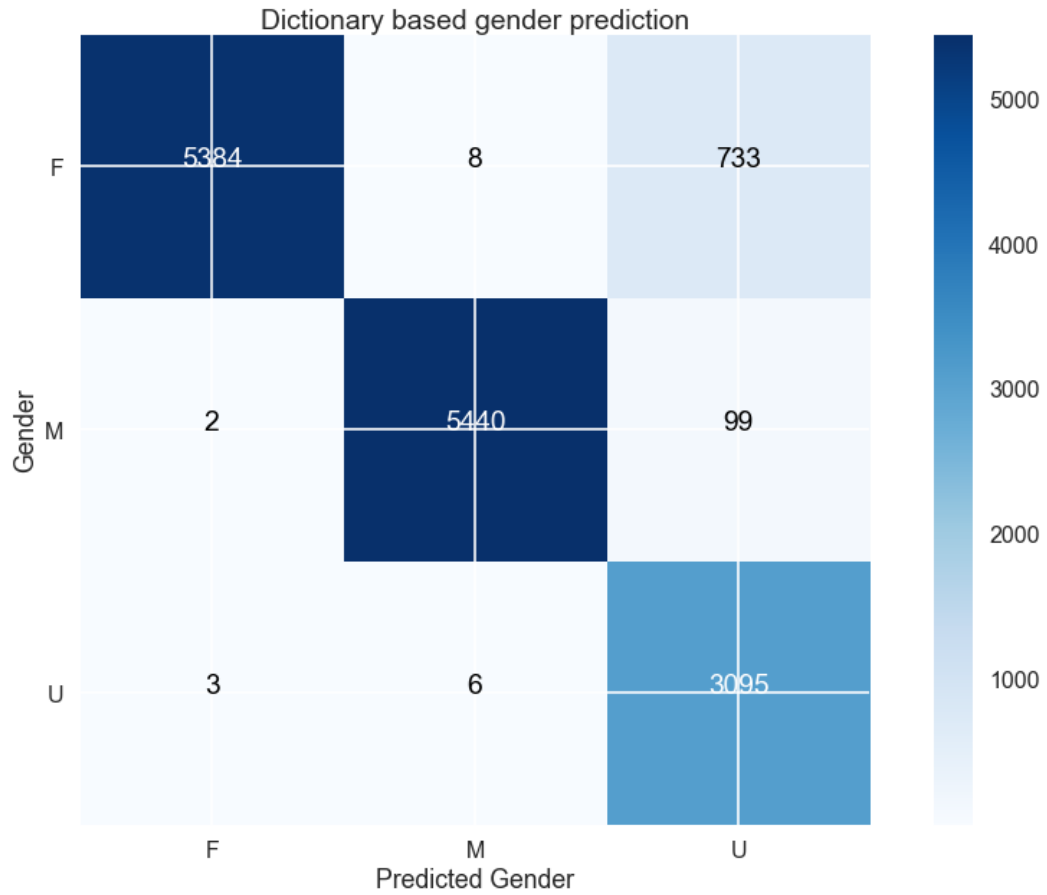

Figure 7. Confusion matrix dictionary-based gender prediction

Figure 7 is the confusion matrix for the dictionary-based gender prediction with Python's generic Gender Guesser.[2] The algorithm is dictionary-based and employs a list of 40,000 given names from all over the world.[3] As can be seen in the confusion matrix, the algorithm predicts known genders very well. The matrix shows the predicted genders in the horizontal axis and the gender in the vertical ones. Each cell describes a combination of gender and predicted gender. For the recorded male and female genders, the algorithm predicts the correct gender. There are only very few inaccuracies such as 8 females that were predicted to be male. Not surprisingly, for the unknown genders there are more errors. The dictionary is based on contemporary names and predicts an unknown gender for 832 vagrants that are either male of female (~21%). The kappa score is 0.911. A kappa statistic considers the possibility of randomly correct predictions, which is something we want to exclude. It measures the observed level of agreement between prediction and actual gender compared to the expected one if the two gender-assignments were totally independent [P.Mean 2008]. A kappa value of 1 indicates a best agreement and 0 the worst. 0.91 can be seen as excellent. The dictionary-based method thus shows great promise, as Blevins and Mullen have also demonstrated [Blevins and Mullen 2015].

In a second dictionary-based approach, we tried to add a historical focus and downloaded the dataset from [Social Security Administration 2017] focussing on historical names. We chose a subset of names that were prominent in the 19th century. As far as we understand Blevins and Mullen, this is the simplified version of their own approach, but using Python instead of R and employing no ensemble method. Results were worse than the generic dictionary-based model. Overall 2,084 gender records were assigned wrongly, which corresponds to a kappa value of 0.79. While this might seem at first disappointing, there are several reasons why the performance might be worse. Either some of the more modern names in [Social Security Administration 2017] were not as modern as we anticipated or, which is also likely for vagrants, some of the names in the datasets did not originate in the English-language world. A dictionary-based approach will not only be dependent on the historical use of words but also on the language used. 37

Overall, however, this does not change the result that the dictionary-based approach delivers excellent results. We could easily address the worse performance in the second approach by adding additional dictionaries, as Blevins and Mullen have done [Blevins and Mullen 2015]. If one is willing to put in the additional work, at some point one will have a perfect combination of dictionaries predicting gender with a very high accuracy. However, this also requires a lot of manual effort creating these dictionaries for various languages and cultures and for different time periods. 38

Let us take a closer look at the detailed results of the first dictionary-based algorithm and any inconsistencies that might reveal interesting relations. Vagrant 7980.1.1 is called “Jesse Paine” and considered female in the dataset. Jesse, however, is historically the father of King David in the Old Testament and therefore the algorithm assigned a male gender. So, this is either a recording mistake or a historically unique Jesse. In total 5 Edwards, Williams and James are also recorded as females, which indicate that these historical male names of English kings have been given to females. On the other hand, vagrant 3849.1 is called “Uphan McGregory”, which the dictionary-based algorithm fails to assign a gender value. 39

Figure 8 presents the most commonly misclassified 'female' names according to the first dictionary-based method. We will discuss the misclassified names later in more detail when we compare the results with other algorithms. 40

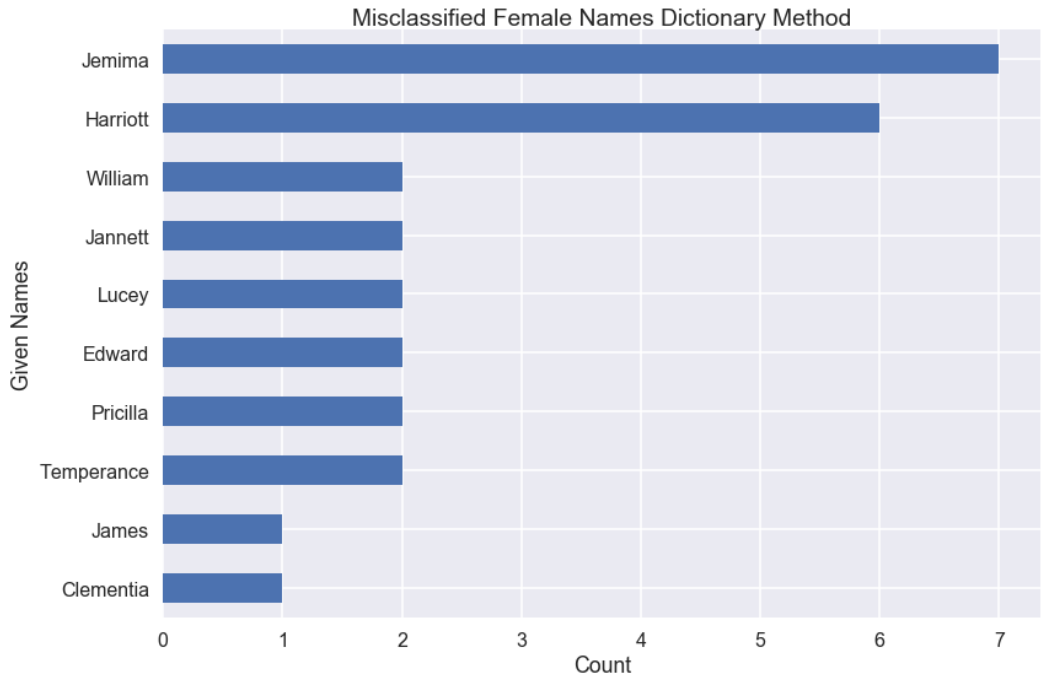

Figure 8. Dictionary-based female misclassifications

Prediction using data-driven techniques: classification and rule-based

In many ways, the dictionary-based approach follows traditional work in the humanities. Computers are faster at looking up items in a dictionary but fundamentally there is no difference here between how a human might approach this problem and a computer. The dictionary approach imitates human techniques to predict the right gender value and is an established method in the digital humanities. Thus, we can use it as a baseline to test predictive analysis of the past with two approaches using machine learning. Regression and rule-based classification are among the best performing models for high-dimensional data according to the empirical evaluation by Caruana, Karampatziakis et al., which is why we have chosen them for this example [Caruana et al. 2008]. Since 2008, neural networks have become an attractive alternative for high-dimensional data, as a more recent evaluation by [Zekić-Sušac et al. 2014] shows. However, the 41

performance difference of neural nets and decision trees was not necessarily significant in their experiments, and neural networks have the disadvantage that they are more difficult to interpret and therefore not immediately useful for a predicting the past approach. Therefore, we will use regression and rule-based classifications. We apply the same dataset first to optimised logistic regression and afterwards to several rule-based systems using rule learners and decision trees. Following our methodology, we first determine the best performing model for each approach using standard predictive analytics steps before we apply this model to the gender analysis of the whole dataset to determine the gender relationships, as the computer sees them.

Classification

In this section, we discuss logistic regression for predicting the past [Raschka 2015]. First all features are included to understand which are the most important ones. Afterwards, various feature selection techniques are used to reduce the dimensionality of the feature space. We also applied a brute-force grid search of all hyperparameters in the model and found the best performance for a L1 regularization with a C=100.0 [Raschka 2015]. 42

Based on a 75-25% training and test split of the vagrant dataset with stratification, we achieved a 76% prediction accuracy on the test dataset for the above configuration of logistic regression: 43

Training accuracy: 0.756
Test accuracy: 0.759

As training and test accuracy are close, the algorithm seems to generalise well. Precision and recall show similar levels of performance in this simple classification model with nearly 0.8 for both: 44

Precision: 0.790
Recall: 0.793

3596 vagrants were misclassified, which leads to a kappa value of 0.62. Despite the good precision and recall values the model significantly underperforms compared to the dictionary-based one. The confusion matrix below shows that logistic regression performed best for unknown genders but confused the assignments of known male and female genders frequently. Let us see why this is the case by investigating which features constitute similarity for logistic regression. 45

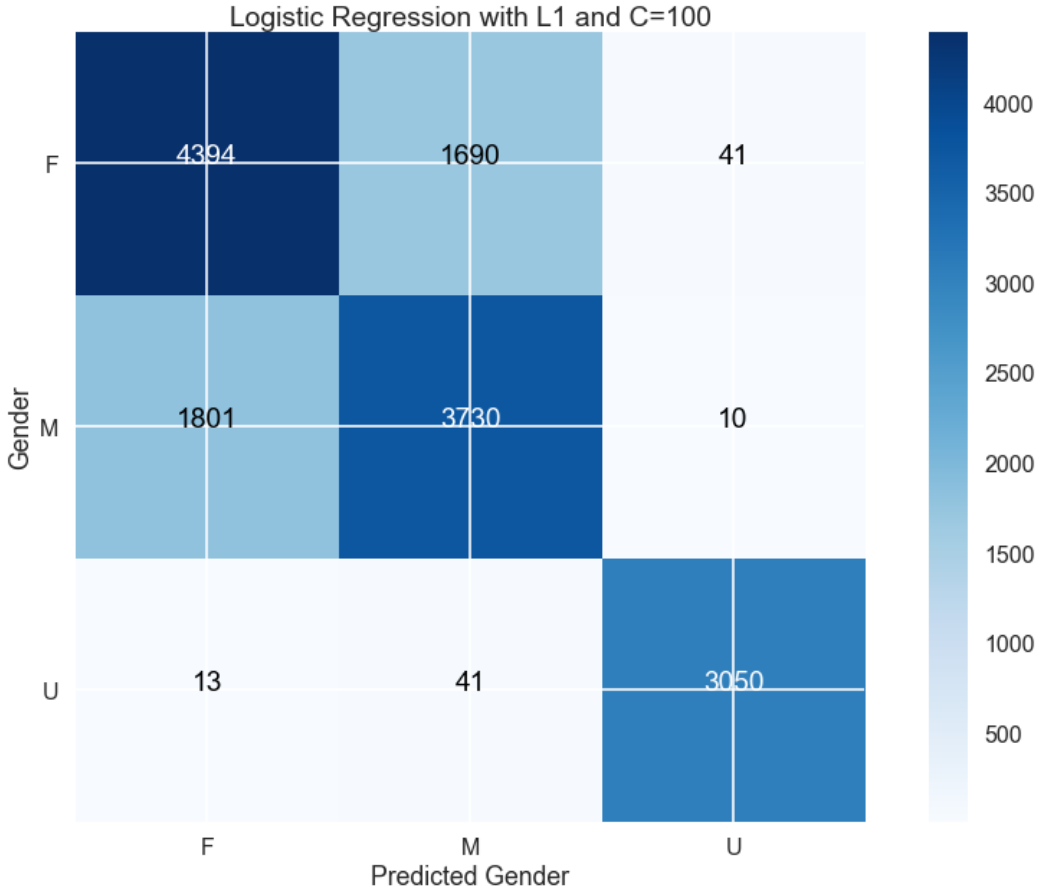

Figure 9. Confusion matrix logistic regression

We apply Sequential Backward Selection (SBS) and Recursive Feature Elimination (RFE) to determine the feature evidence, which contributes most to determining the gender [Raschka 2015]. Both algorithms iteratively remove the worst-performing features until an optimum minimal set is reached. The selected top for SBS are: “Relationship to Lead Vagrant”, “Settlement Country”, “Given Names”, “Number of People in Group” and “SessionStart”. For RFE, the following top 5 features were selected: “Number of People in Group”, “SessionStart”, “Given Names”, “Relationship to Lead Vagrant” and “Settlement Country”. This means both algorithms agree on the most important features that determine a vagrant's gender. We immediately notice that “Given Names” - so important for the dictionary-based method - seems to contribute less to the gender decision than other features for both SBS and RFE, which makes us suspect that either there is a strong unknown relationship in the data or logistic regression does not optimise the selection well. 46

The following table compares the performance of the full feature set with the top 5 features selected by SBS and RFE respectively: 47

Model	Training Accuracy	Test Accuracy
All features	0.756	0.759
Top features with SBS	0.756	0.761
Top features with RFE	0.752	0.752

Table 2.

The good news therefore is that the reduction of the number of features and thus a reduction of the problem of sparse data does not significantly influence the performance of the classification. Overall, however, the performance of this classification algorithm is not satisfactory in comparison to the dictionary-based one. It falls short when trying to decide the difference between male and female genders compared with the traditional approach to predicting genders based on first names. As the feature selection discussion demonstrates the algorithm is not able to determine the most important feature, which is the given name. 48

Figure 10 is the learning curve of logistic regression, which shows that the performance does not improve with additional training data. 49

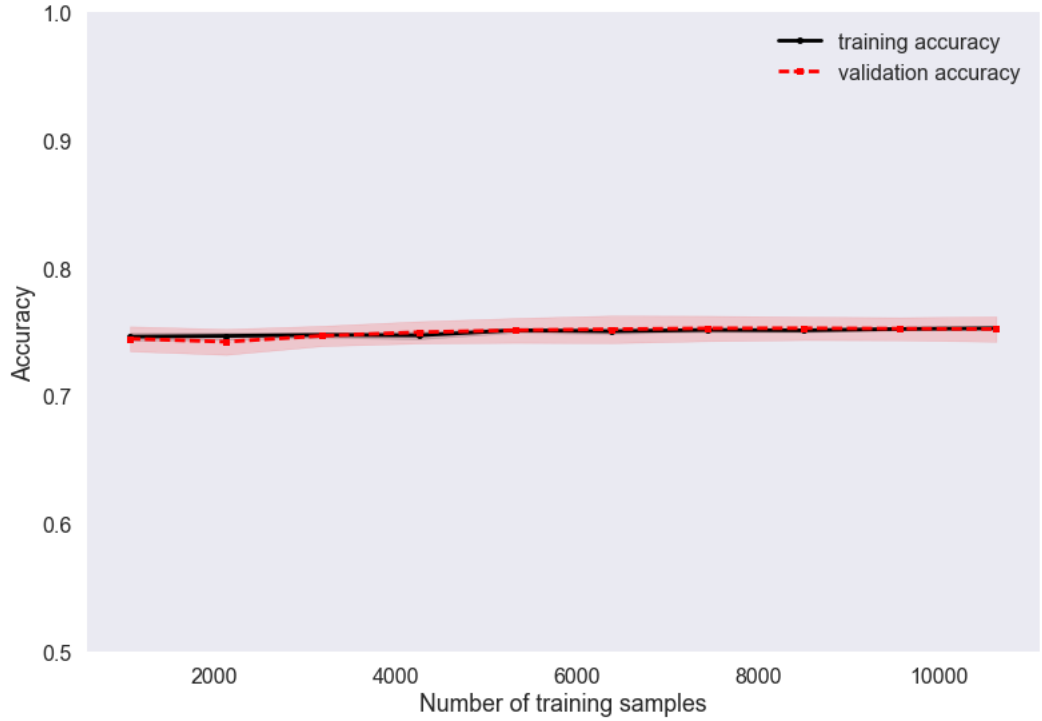

Figure 10. Learning curve logistic regression

Learning curves help us determine whether we have enough data for the number of features. Here, it is flat, which means it is not so much the lack of data but issues with the approach, which leads to the comparably worse performance. A standard classification approach is not good enough for predicting the past of genders even if we add more training samples. 50

Rule-based algorithms

Next in our series of experiments with predictive analytics are so-called rule learners – the second class of best performing models. These represent knowledge as a set of rules or logical if-else statements; described by the antecedent and the consequence [Lantz 2013, 142]. So, a simple rule in our case would be “If the given name is Anne then it is a female”. For machine learning, this means that the if-statement consists of a logical combination of features (predictors), while the result is a decision on the gender of the vagrant under observation (target). We could therefore use the antecedent to understand the conditions and characteristics that define decisions on vagrants – in our case in terms of gender. These rule learners have the additional advantage that they behave well with nominal features and identify uncommon observations [Lantz 2013, 142]. Finally, they are relatively easy to interpret. All of these characteristics should make rule learners interesting to digital humanities. 51

All rule-based learners work by splitting up the data into smaller and smaller groups, which finally can be formulated by a particular rule. The first algorithm we tested was Weka's OneRule [Holte 1993]. The Waikato Environment for Knowledge Analysis (Weka) is a machine learning platform, developed at the University of Waikato, New Zealand.[4] OneR (short for OneRule) generates one rule for each predictor, then selects the rule with the best performance towards the target as the “one rule”. We ran this directly in Weka, as we did not have to reuse the results. We only present the first couple of rows of the output: 52

Given Names: 53

< 1.5 -> F
< 4.5 -> M
< 5.5 -> F

This output might look strange at first sight, but it simply says that the vagrants below the second given name is female (indexes in alphabetic order smaller than 1.5), then everything until the fifth name is male, then we find another female and so on. OneR thus learned without prior knowledge that genders are best determined by the given names. This is not surprising as it operates similarly to the method described in Blevins and Mullen 2015 but by learning the relationship from the dataset [Blevins and Mullen 2015]. If the dataset contains given names with several gender 54

associations, OneR considers the most commonly appearing one as the right decision. Please note that the datasets were shuffled before applying OneR to randomize the input. It is only the output of OneR that is sorted.

55 The OneR learner also outperforms the classification algorithm with a kappa statistic of 0.98 and is thus better than the traditional method using the dictionaries. While OneR therefore is excellent at deriving gender by mainly outperforming the traditional approaches in terms of the unknown gender assignments, we do not learn very much from the rules it has come up with. But compared to logistic regression, the OneR learner has identified the most important feature to learn about vagrants' genders without any background knowledge.

56 The Ripper algorithm [Cohen 1995] in Weka is an improvement on the OneR learners, as it can consider multiple features. Ripper is based on a complex set of heuristics. We can therefore hope to learn more from it about our data's composition. The Weka version of Ripper learns overall 40 rules from our data. Examples include:

- '(Given Names >= 435) => Gender=[unknown]' is similar to the OneR algorithm result in so far as the rule is using the single most important feature and determines given names with an index larger-equal to 435 as of unknown gender. These values include '[illegible]', ' [none given]' or '[omitted]'.
- However, the algorithm also learns the rule that should the observed person be a child of a lead vagrant and we do not know anything else about it, its gender should be unknown: Relationship to Lead Vagrant <= 0 => Gender=[unknown]. 0 is the label the child relationship in that feature.
- Finally, Ripper delivers more complex rules like (Settlement Country >= 7) and (Number of People in Group <= 1) and (Given Names >= 182) and (Given Names <= 196) => Gender=M, which describe special cases where in addition to the given name (in this case Harold, Henry, etc.) we have further information to determine the gender such as that the settlement country is Barbados.

57 Again, kappa is 0.98 for Ripper, which shows that the algorithm performs very well. We have, however, also learned additional characteristics that decide the genders of vagrants and could use in a historical analysis to investigate inconsistencies further. Is there, e.g., a reason why Barbados plays a role in the gender of vagrants or is this just noise?

58 Next, we work through a decision tree analysis of the Vagrant data, which is the most popular rule-based machine learner [Raschka 2015]. In a decision tree, we follow the path from the root of the tree to its leaves. Each turn describes one decision or rule. In terms of deriving gender, suppose we would (hypothetically) say that the first (root) decision would be whether vagrants were judged by Magistrate Clark. If that were the case, the next decision were whether the session took place before June. If both would be the case, we would assume that the vagrant was female and otherwise we need to follow a different path. Each time we make a decision based on one feature and then work ourselves down the tree to come to a conclusion. The decision tree can learn therefore a much more complicated system of rules than the previous two rule learners.

59 Next we work through our predicting the past methodology again, starting with the standard predictive analytics steps. Brute-force testing the hyperparameters first, we can see a very good performance as we increase the depth but also the danger of overfitting the tree to the particular dataset. We use a min_sample_split of 20 to control overfitting.

60 We train the tree on a 75%-25% training and test split with stratification. The learned model is then used to predict all genders. The result is the confusion matrix in Figure 11, which presents an excellent performance. The algorithm only misses out on 187 vagrants (or about 1% of the whole). Compared to the dictionary-based method, wrong predictions are relatively evenly distributed among male, female and unknown genders but there are more mistakes for known males and females.

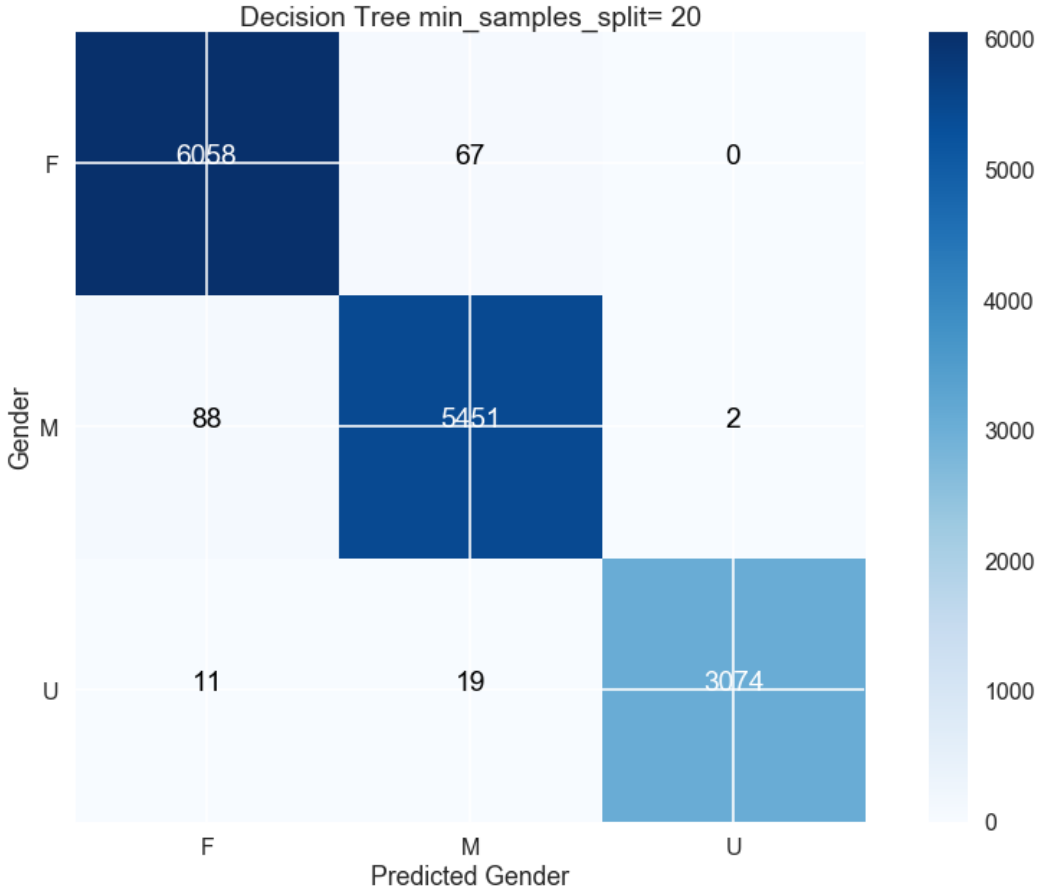

Figure 11. Confusion matrix decision tree

Investigating feature importance next, we generalise the decision tree to a RandomForest ensemble [Raschka 2015] to avoid overfitting. We can clearly see in Figure 12 how much each of the features contributes to the gender prediction of each vagrant and how well the ensemble identifies the most important features. 61

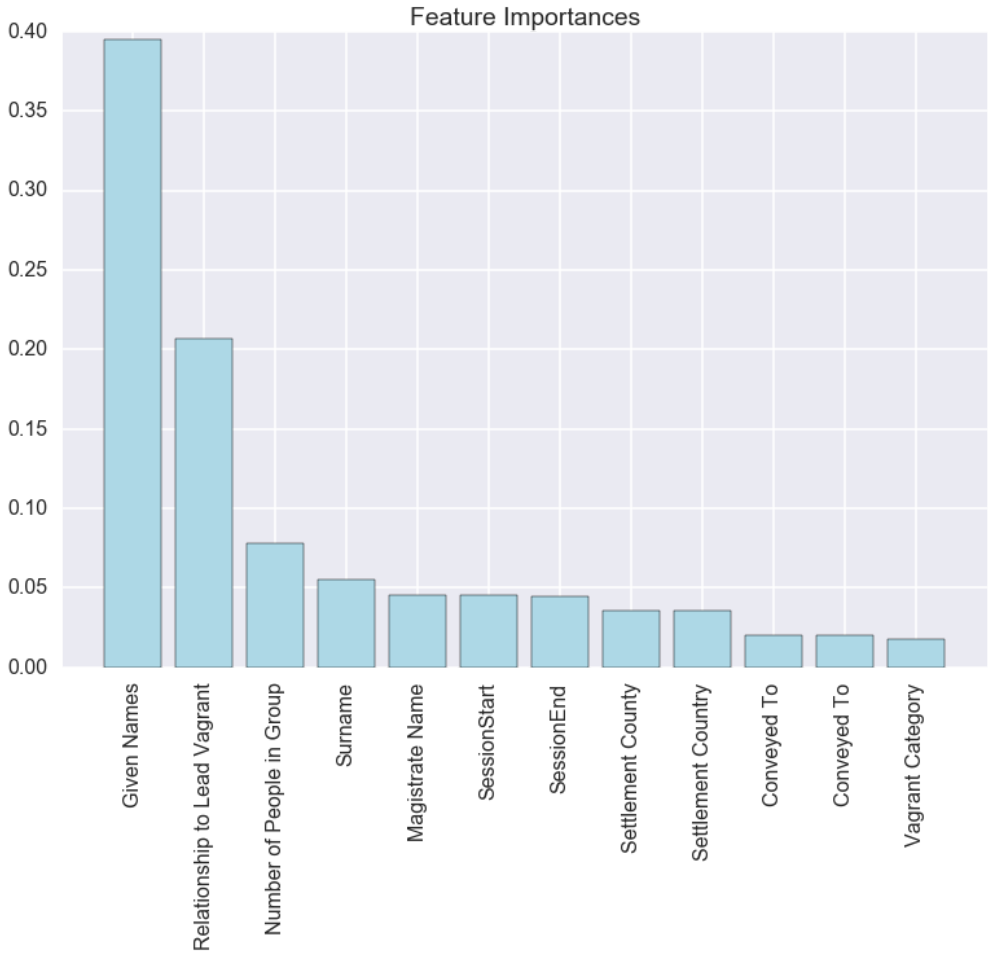

Figure 12. Feature importance random forest

With these results, we can conclude that names, relationships as well as the number of people in a group as a proxy to the family status determine gender, while magistrate, timing as well as places seem to have had a negligible influence. This means that the magistrates did not judge per gender and that there is also no season for males or females. Overall this result from the RandomForest largely confirms what we found for the SBS as well as RFE algorithms above, but this time given names are clearly the most important feature. 62

Figure 13 is the decision tree learning curve, which shows good performance and small improvements with more training samples. 63

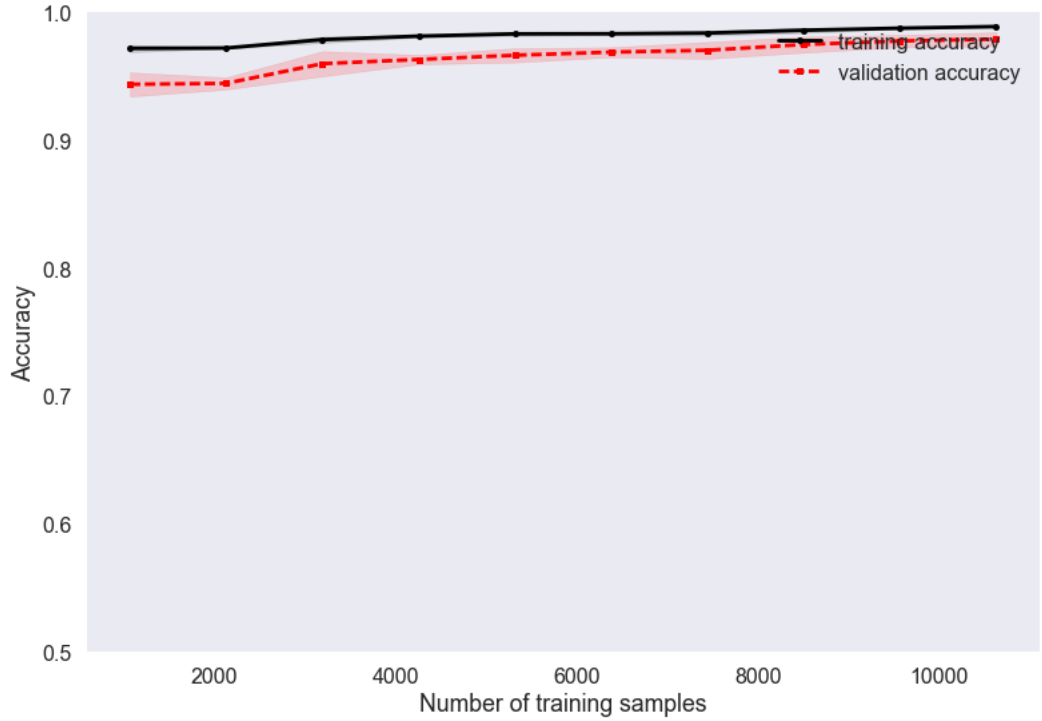

Figure 13. Learning curve decision tree

64 The kappa score is 0.98, similar to the other rule learners and much better than the dictionary-based models. Compared to these, the decision tree is much better at predicting unknown female and male genders but worse at predicting known female and male genders. Decision trees are obviously able to learn unknown relationships in the data, which the dictionary does not contain. Our decision tree has learned to interpret “[Wife]” as a female given name. For instance, for vagrant 5999.2.2, the name is recorded as “[Wife] Young” and the gender as female. While the dictionary-based method does not assign a value, the decision tree has learned that this is a female. The dictionary-based method on the other hand leads to a large number instances like this, where an unknown gender was predicted while the actual gender was female. This could of course be easily corrected in the dictionary-based algorithm by adding a dedicated rule for “[Wife]” but it does demonstrate that decision trees have learned from the available data. The dictionary-based gender assessment, on the other hand, relies on existing background knowledge and theories.

65 Compared to the dictionary-based method, the decision tree has learned that vagrant 10436.1.3 called Auquin Jeffs is a male, which was unknown to the dictionary. On the other hand, the decision tree did not learn that Ave Smith is a male. Possibly because there is only one vagrant named Ave in the whole dataset and all the given names in its tree-neighbourhood like Archibald, Arnold, Arthur, etc. were male. This led to Ave being added to a male leaf in the decision tree. Both dictionary-based method and decision tree struggled with typical male names such as Edward, which were recorded for female vagrants like vagrant 10018.1.1. The decision tree assigns a female gender in only one such case: vagrant 382.1.2 called William Bourne. But at least the decision tree learned this specific case compared to the dictionary-based method. Finally, there are many cases where a decision tree had learned the same gender for a given name as the one that is recorded in the dictionary but the record showed “unknown” gender. Vagrant 6396.1.1 called Din Foley is an example.

66 Compared to the dictionary-based methods the decision tree relatively evenly misses out on either female or male genders. A comparison of the top 20 female names, the decision tree misclassifies, can be seen in Figure 14.

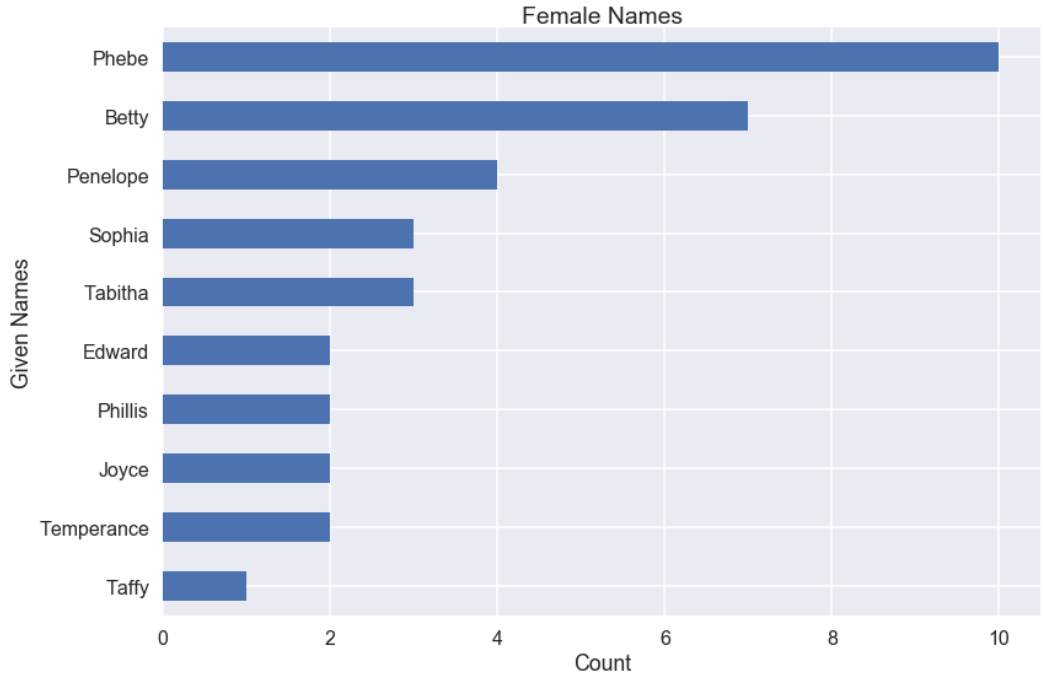

Figure 14. Decision tree female misclassifications

As seen above in Figure 8, Jemima as well as Harriott seem to be not known to the dictionary-based method. The decision-tree prediction model, on the other hand, predicts these correctly as female. With predictive analytics, we have thus achieved a new perspective on the gender target that allows for insights going beyond established additional knowledge as encoded in the dictionary. On the other hand, the decision tree misses out on Phebe and Betty, which could be female names. It does not have enough training data to know them. 67

The dictionary-based method is better at discriminating those vagrants where the gender is known to be either female or male. It hardly makes any mistakes here. But it fails to deal with any unknown relationship such as the fact that “[wife]” is female. Compared to the other rule learners the decision tree rules are much more complex. Looking at the rules the decision tree learned, the root of the tree assigns an unknown gender to all children, whose relationship to the lead vagrant is not specified. This is similar to the second rule cited above for the Ripper learner. Then, the decision tree mainly uses a long chain of conditions on the feature “Given Names” to walk down the tree to assign the gender. But contrary to the simple rule learner, it also employs other features to control the gender. So, for instance, there is a classification of vagrants as female based on their given name, unless they were dealt with by magistrate “Benj Cheny” and some of his related colleagues. 68

In conclusion, decision trees (and other rule-based learners) work where existing theories on the relationship between a given name and gender contradict the underlying information, because the rule-based learners are working close to the data. Here, predictive analytics targeting the past opens cases that should be relevant to the digital humanities, as data contradicts existing historical ideas. From a conceptual point of view, we could develop new ideas about the use of female names; thereby also probably improving the theory of the link between given names and genders. In terms of data practices, we could use our predictive insights to enhance the dictionary of names. Shmueli speaks of the possibility through prediction to provide a “reality-check” for theories and how far they differ from what can be observed [Shmueli 2010]. In our case, we reality-check how modern theories of names' genders correspond to past ones. The approach also works for specific societies and language spaces, for which we have no dictionaries. Finally, prediction allows to understand its own limits by drawing attention to the fact that not all things in life can be predicted. Not all “James” are male, which decision trees can tell us by going through a detailed analysis of their errors. 69

Conclusion

In this paper, we discussed a digital humanities approach inspired by predictive analytics, which we called predicting the past. We introduced the background of this approach and recent discussions in predictive analytics that develop it to investigate existing methods against integrated new datasets. Afterwards we performed a study how to predict past genders using the Vagrant Lives dataset. The high-dimensionality of the data, typical to the humanities, makes it especially suitable for a predictive investigation. 70

71

Predictive analytics need models to demonstrate relationships between the features we investigate. These models are learned against the training data. Otherwise, gender prediction in the digital humanities uses predefined gender dictionaries of first names and then matches the gender of individuals against this dictionary. This has firstly the problem that these dictionaries are heavily dependent on culture and language they relate to. But this is not the only issue, as dictionary-based approaches secondly also assume that errors are randomly distributed. Gender trouble is simply a problem of not recording the right gender in the data. Our predictive analytics approach on the other hand does not make this assumption in advance and judges gender purely based on the existing data. In our case, this meant that unknown genders are kept unknown if there is not enough evidence to the contrary and even if the first name was seemingly about a particular gender. Maybe the original assignment of an unknown gender to James was not an error but rather an indicator that the gender was difficult to determine.

The dictionary-based approach is an accepted methodology in the digital humanities to derive genders in datasets. It could thus provide us with a baseline to compare the performance of advanced machine learning approaches for predictive analytics. In the words of Schutt and O'Neil, we used the dictionary-based approach to fine-tune our “spam filter” of models and understand what performance level predicting the past models had to reach to be taken seriously for gender prediction [Schutt and O'Neil 2013]. We considered only predictive analytics models known to perform well in supervised machine-learning and experimented first with classification algorithms, which did not have a satisfactory performance. Afterwards, we used rule-based learners, which matched the dictionary-based models and sometimes surpassed their performance, while they did not have the requirement that dictionaries are created beforehand. They learned how to learn about genders from the data in all its categories of male, female and unknown. 72

Dictionary-based methods remain clearly one of the best ways to determine the correct genders if we suspect that there are many recording and data entry mistakes for these. Correcting genders then can lead to more consistent follow-on investigations, as Blevins and Mullen have demonstrated [Blevins and Mullen 2015]. But they are only possible if we have high-quality records of given names and their gender, not just for a particular language but also for a certain culture and time period. If that is not the case, this paper has demonstrated that we have machine learning alternatives at hand from the world of predictive analytics that show an equivalent performance. 73

If we are concerned not just with determining the correct gender but understanding the rules that governed gender composition in a dataset then predictive analytics approaches can offer genuine new insights. Predicting the past can be an interesting toolkit for the digital humanities; working with non-textual data just like distant-reading works for texts. It depends, however, on the existence of high-quality datasets such as Vagrant Lives and the ability to integrate them. With the automation of machine learning processing and the availability of Weka, R and Python toolkits to interface with advanced machine learning algorithms, predicting the past can become part of a future day-to-day research in digital humanities. It does currently depend on some programming knowledge but far less so than it did only a couple of years ago. In the future, technical knowledge will count less than the ability to read the results of machine learning processes and to situate them within their data. All this makes predicting the past an exciting new approach to explore in the digital humanities. 74

Notes

[1] http://openhumanitiesdata.metajnl.com/

[2] https://pypi.python.org/pypi/gender-guesser/

[3] https://www.heise.de/ct/ftp/07/17/182/

[4] https://www.cs.waikato.ac.nz/ml/weka/

Works Cited

Abbott 2014 Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*, John Wiley & Sons.

Anadiotis 2016 Anadiotis, G. (2016). “Predictive analytics and machine learning: A dynamic duo.” Available at http://www.zdnet.com/article/predictive-analytics-machine-learning/. [Accessed August 26, 2017].

Anderson and Blanke 2012 Anderson, S. and T. Blanke (2012). “Taking the long view: from e-science humanities to humanities digital ecosystems.” *Historical Social Research/Historische Sozialforschung* 37(3): 147-164.

Anderson et al. 2010 Anderson, S., T. Blanke and S. Dunn (2010). “Methodological commons: arts and humanities e-Science fundamentals.” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368(1925): 3779-3796.

Aradau and Blanke 2016 Aradau, C. and T. Blanke (2016). “Politics of prediction Security and the time/space of governmentality in the age of big data.” *European Journal of Social Theory* 20(3): 373-391.

Argamon and Olsen 2009 Argamon, S. and M. Olsen (2009). “Words, patterns and documents: experiments in machine learning and text analysis.” *Digital Humanities Quarterly* 3(2).

Blanke and Hedges 2013 Blanke, T. and M. Hedges (2013). “Scholarly primitives: Building institutional infrastructure for humanities e-Science.” *Future Generation Computer Systems* 29(2): 654-661.

Blevins and Mullen 2015 Blevins, C. and L. Mullen (2015). “Jane, John... Leslie? a historical method for algorithmic gender prediction.” *Digital Humanities Quarterly* 9(3).

Bowles 2015 Bowles, M. (2015). *Machine learning in Python: essential techniques for predictive analysis*, John Wiley & Sons.

British Library 2015 British Library. (2015). “The British Library Machine Learning Experiment.” Available at http://blogs.bl.uk/digital-scholarship/2015/04/the-british-library-machine-learning-experiment.html. [Accessed August 26, 2017].

Caruana et al. 2008 Caruana, R., N. Karampatziakis and A. Yessenalina (2008). *An empirical evaluation of supervised learning in high dimensions*. Proceedings of the 25th international conference on Machine learning, ACM.

Choi and Varian 2012 Choi, H. and H. Varian (2012). “Predicting the present with google trends.” *Economic Record* 88(s1): 2-9.

Cohen 1995 Cohen, W. W. (1995). *Fast effective rule induction*. Proceedings of the twelfth international conference on machine learning.

Crymble et al. 2015 Crymble, A., L. Falcini and T. Hitchcock (2015). “Vagrant lives: 14,789 vagrants processed by the county of Middlesex, 1777–1786.” *Journal of Open Humanities Data* 1.

Firat 1987 Firat, A. F. (1987). “Historiography, scientific method, and exceptional historical events.”, in Melanie Wallendorf and Paul Anderson (eds.), *Advances in Consumer Research*, Vol. 14, pp. 435–8. Provo, UT: Association for Consumer Research.

Gaffney 2008 Gaffney, C. (2008). “Detecting trends in the prediction of the buried past: a review of geophysical techniques in archaeology.” *Archaeometry* 50(2): 313-336.

Grimmer and Stewart 2013 Grimmer, J. and B. M. Stewart (2013). “Text as data: The promise and pitfalls of automatic content analysis methods for political texts.” *Political analysis* **21**(3): 267-297.

Holte 1993 Holte, R. C. (1993). “Very simple classification rules perform well on most commonly used datasets.” *Machine learning* 11(1): 63-90.

Jasinski 2017 Jasinski, M. E. (2017). “Predicting the Past — Materiality of Nazi and Post-Nazi Camps: A Norwegian Perspective.” *International Journal of Historical Archaeology*, 2017: 1-23.

Jockers and Witten 2010 Jockers, M. L. and D. M. Witten (2010). “A comparative study of machine learning methods for authorship attribution.” *Literary and Linguistic Computing* 25(2): 215-223.

Juola 2008 Juola, P. (2008). “Authorship Attribution.” *Foundations and Trends in Information Retrieval* 1(3): 233-334.

Jänicke et al. 2015 Jänicke, S., G. Franzini, M. F. Cheema and G. Scheuermann (2015). “On close and distant reading in digital humanities: A survey and future challenges.” Eurographics Conference on Visualization (EuroVis)-STARs. The Eurographics Association.

Katz et al. 2014 Katz, D. M., I. Bommarito, J. Michael and J. Blackman (2014). “Predicting the behavior of the supreme court of the united states: A general approach.” *arXiv preprint arXiv:1407.6333*.

Klingenstein et al. 2014 Klingenstein, S., T. Hitchcock and S. DeDeo (2014). “The civilizing process in London’s Old Bailey.” *Proceedings of the National Academy of Sciences* 111(26): 9419-9424.

Lantz 2013 Lantz, B. (2013). *Machine learning with R*, Packt Publishing Ltd.

Lincoln 2017 Lincoln, M. (2017). “Predicting the Past: Digital Art History, Modeling, and Machine Learning.” Available at http://blogs.getty.edu/iris/predicting-the-past-digital-art-history-modeling-and-machine-learning/. [Accessed March 1, 2018]

Manovich 2015 Manovich, L. (2015). “Data Science and Digital Art History.” *International Journal for Digital Art History*, 1(1).

P.Mean 2008 P.Mean. (2008). “What is a Kappa coefficient? (Cohen's Kappa).” Available at http://www.pmean.com/. [Accessed March 1, 2018]

Popper 1974 Popper, K. R. (1974). “Prediction and prophecy in the social sciences.”" In Karl Popper: *Conjectures and Refutations.* London: Rutledge and Kegan Paul.

Raschka 2015 Raschka, S. (2015). *Python machine learning*, Packt Publishing Ltd.

Röhle 2012 Röhle, B. R. T. (2012). “Digital Methods: Five Challenges.” in *Understanding Digital Humanities*. D. M. Berry. London, Palgrave Macmillan UK: 67-84.

Schutt and O'Neil 2013 Schutt, R. and C. O'Neil (2013). *Doing data science: Straight talk from the frontline*, O'Reilly Media, Inc.

Shmueli 2010 Shmueli, G. (2010). “To explain or to predict?” *Statistical science* 25(3): 289-310.

Social Security Administration 2017 Social Security Administration. (2017). “Get Ready For Baby.” Available at https://www.ssa.gov/oact/babynames/index.html. [Accessed August 26, 2017]

Stanley 2012 Stanley, M. (2012). “Predicting the Past: Ancient Eclipses and Airy, Newcomb, and Huxley on the Authority of Science.” *Isis* 103(2): 254-277.

Tench et al. 2016 Tench, S., H. Fry and P. Gill (2016). “Spatio-temporal patterns of IED usage by the Provisional Irish Republican Army.” *European Journal of Applied Mathematics* 27(03): 377-402.

Wang et al. 2013 Wang, T., C. Rudin, D. Wagner and R. Sevieri (2013). “Learning to Detect Patterns of Crime.” *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*. H. Blockeel, K. Kersting, S. Nijssen and F. Železný. Berlin, Heidelberg, Springer Berlin Heidelberg: 515-530.

Weisskopf 1984 Weisskopf, V. F. (1984). “The Frontiers and Limits of Science.” *Daedalus* 113(3): 177-195.

Zekić-Sušac et al. 2014 Zekić-Sušac, M., S. Pfeifer and N. Šarlija (2014). “A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem.” *Business Systems Research Journal*. 5: 82.